"In presenting the dissertation as a partial fulfillment of the requirements for an advanced degree from the Georgia Institute of Technology, I agree that the Library of the Institution shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to copy from, or to publish from, this dissertation may be granted by the professor under whose direction it was written, or, in his absence, by the dean of the Graduate Division when such copying or publication is solely for scholarly purposes and does not involve potential financial gain. It is understood that any copying from, or publication of, this dissertation which involves potential financial gain will not be allowed without written permission.

_____ "

EXACT PROBABILITIES OF THE KOLMOGOROV - SMIRNOV STATISTIC

A THESIS

Presented to

the Faculty of the Graduate Division

by

Oscar Vernon Hefner

In Partial Fulfillment

of the Requirements for the Degree

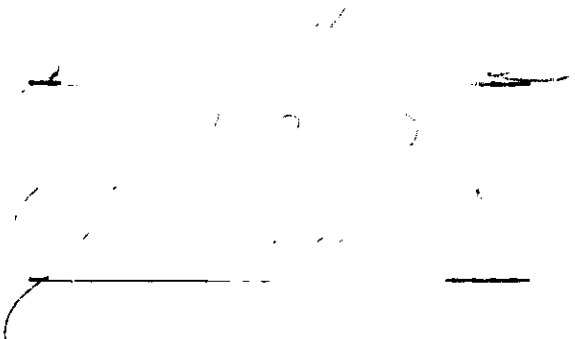
Master of Science in Applied Mathematics

Georgia Institute of Technology

June, 1960

EXACT PROBABILITIES OF THE KOLMOGOROV - SMIRNOV STATISTIC

Approved:



Date approved by Chairman:

23 May 1960

ACKNOWLEDGEMENTS

The author would like to express his sincere appreciation to his advisor, Dr. James W. Walker, for the suggestion of the topic of this study and for his guidance and encouragement in its preparation.

The author would further like to express his appreciation for the able assistance of the staffs of the Rich Electronic Computer Center at the Georgia Institute of Technology and the Research Computation Center at the University of North Carolina.

TABLE OF CONTENTS

LIST OF TABLES

SUMMARY

Let X be a random variable with a probability distribution described by the continuous cumulative distribution function

$$F(x) = P(X \leq x) \quad .$$

Next, let $(x_1, x_2, \ldots, x_n)$ be n independent observations of X, written in increasing order

$$x_1 \leq x_2 \leq \ldots \leq x_n \quad .$$

The empirical cumulative distribution function, $F_n(x)$, is defined by

$$F_n(x) = \frac{1}{n}(\text{number of } x_i \text{ such that } x_i \leq x) \quad .$$

$F_n(x)$ is a step function with jumps at

$$x = x_1, x_2, \ldots, x_n \quad .$$

From the law of large numbers one concludes that, for large n, the probability that $F_n(x)$ will differ from F(x) by very little is very close to one. Consider

$$D_n = \operatorname{lub}_x \left| F(x) - F_n(x) \right| ,$$

the greatest absolute discrepancy between $F(x)$ and $F_n(x)$. It follows that D_n itself is a random variable and, again from the law of large numbers, one concludes that for n large the probability of D_n being very small is very close to one. Kolmogorov improved this statement by proving that:

(a) D_n is, for fixed sample size n, a random variable with a probability distribution which is independent of the probability distribution $F(x)$ of the original random variable X, and

(b) for $z > 0$, we have uniformly in z

$$\lim_{n\to\infty} P\{\sqrt{n}\, D_n \leq z\} = 1-2\{\sum_{v=1}^{\infty} (-1)^{v-1} e^{-2v^2z^2}\} = L(z) .$$

This last statement gives the probability distribution of D_n for large n.

The use of these results in practical applications is hindered by the fact that the sample size may be small, and the question arises as to the error introduced by using the limiting distribution. To overcome this difficulty one could derive a bound on the error or find a means by which the exact probability for finite n could be determined. Various studies in both of these areas have been made.

The present study is concerned with the derivation of certain recursive relations by means of which the exact distribution function of D_n can be calculated for rational numbers with denominator n. More precisely, a careful analysis of W. Feller's development of Kolmogorov s results leads to the relations

$$u_1 = s_1$$

$$u_k = s_k - \sum_{r=1}^{k-1} u_r a_{kr} \qquad (k = 2,\ldots,c)$$

$$u_k = s_k - \sum_{r=1}^{k-1} u_r a_{kr} - \sum_{r=c}^{k-1} v_r b_{kr} \qquad (k = c+1,\ldots,n-c)$$

$$v_c = t_c$$

$$v_k = t_k - \sum_{r=c}^{k-1} v_r d_{kr} \qquad (k = c+1,\ldots,2c)$$

$$v_k = t_k - \sum_{r=c}^{k-1} v_r d_{kr} - \sum_{r=1}^{k-2c} u_r c_{kr} \qquad (k = 2c+1,\ldots,n-1)$$

and

$$P\{nD_n > c\} = \sum_{r=1}^{n-c} u_r + \sum_{r=c}^{n-1} v_r$$

where a_{kr}, b_{kr}, c_{kr}, d_{kr}, s_k, t_k are certain calculable binomial probabilities.

Tables compiled from these relationships using the IBM 650 Magnetic Drum Data-Processing Machine and the Remington Rand Univac Scientific 1105 Electronic Computer are presented at the end of the study.

CHAPTER I

INTRODUCTION

Let $F(x)$ denote the common cumulative distribution function (cdf) of the mutually independent random variables $(X_1, X_2, \ldots, X_n)$; and let $(x_1, x_2, \ldots, x_n)$ denote a sample taken on $(X_1, X_2, \ldots, X_n)$ arranged in order of magnitude. That is,

$$x_1 \leqq x_2 \leqq \ldots \leqq x_{n-1} \leqq x_n \quad .$$

The empirical distribution of the ordered sample $(x_1, x_2, \ldots, x_n)$ is the step function $F_n(x)$ defined by

$$F_n(x) = \begin{cases} 0 & \text{for } x < x_1 \quad , \\ \frac{k}{n} & \text{for } x_k \leqq x < x_{k+1} \quad , \\ 1 & \text{for } x \geqq x_n \quad . \end{cases}$$

The deviation of $F_n(x)$ from $F(x)$ is of interest in various problems in the theory of non-parametric estimation, and in this direction certain investigations have been made on the probable deviations of $F_n(x)$ from $F(x)$. In particular, the ω^2-criterion of von Mises [1] gives estimates for the probable deviations for certain forms of $F(x)$; and more

general results have been given by Kolmogorov [2] and Smirnov [3] for F(x) any continuous cdf. In the latter two investigations, as the sample size tends to infinity, a limit distribution of the probable deviations has been found. The present study is primarily concerned with the approach taken by Kolmogorov.

Consider the random variable D_n defined by

$$D_n = \operatorname*{lub}_x \left| F_n(x) - F(x) \right| \quad .$$

Although the exact distribution of D_n is not known, Kolmogorov showed that $\sqrt{n}\, D_n$ has a limiting distribution. More precisely, we have

Theorem I (Kolmogorov [4]). Suppose that F(x) is continuous, then for every z as $n \to \infty$

$$P\{\sqrt{n}\, D_n \leqq z\} \to L(z)$$

where

$$L(z) = 1 - 2 \sum_{v=1}^{\infty} (-1)^{v-1} e^{-2v^2 z^2} \qquad \text{for } z \geqq 0 ,$$

$$0 \qquad \text{for } z < 0 .$$

Kolmogorov's proof to Theorem I is lengthy, as is a proof of the theorem due to Smirnov [5]. Simpler proofs

have been given by Feller [6] and Doob [7].

Application of Theorem I to problems involving finite sample sizes must be made with some reservation in terms of the error introduced by the use of the limiting distribution. A bound on this error is given by Chung [8] in the following theorem: (it is stated in its original form, although some of the expressions differ in notation from those previously cited.)

Theorem 2 (Chung). Let

$$\phi(\lambda) = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2\lambda^2}$$

and

$$d_n = \sup_x \left|F_n(x) - F(x)\right| \quad .$$

If $A_o > 0$ is an arbitrary constant and if $\lambda(n)$ is a function of n such that

$$\frac{1}{A_o \log n} \leq \lambda(n) \leq A_o \log n \quad ,$$

then

$$\left| P\{d_n \leqq \sqrt{n}\ \lambda(n)\} - \phi\left[\lambda(n)\right] \right| \quad =$$

$$A\, n^{-1/10} \left[1 - \sqrt{\log n}\ \left(\lambda^{-1}(n) + \lambda^{-2}(n)\right)\right]$$

where A is a constant depending only on A_o.

The emphasis in this study is on the investigation of a method by which $P\{nD_n \leqq c\}$ can be computed for finite n. Kolmogorov [4], gives recursive relations for calculating $P\{nD_n \leqq c\}$ for finite n, and his method has been modified by Massey [9] and later by Birnbaum [10]. These relations follow from considering the probability that the step function $F_n(x)$ will lie entirely inside the band

$$F(x) - \frac{c}{n} < F_n(x) < F(x) + \frac{c}{n} \quad .$$

By a geometric argument it can be shown that the <u>a priori</u> probability of $F_n(x)$ being inside the band has a certain multinomial distribution. It is from this distribution that a set of recursive relations can be formed. Massey extended this argument to derive a system of 2c - 1 linear homogeneous difference equations which yield the desired probabilities.

In Chapter II, recursive relations different from those of Kolmogorov are developed. The development of these relations is accomplished by a careful study of the method of Feller in proving Kolmogorov's theorem. In his paper, Feller derives a system of 2n linear equations, the solution of which gives the desired exact probability. It is shown here

that these equations reduce to simple recursive relations if the definitions and restrictions of the quantities involved are carefully considered throughout the argument.

CHAPTER II

RECURSIVE RELATIONS FOR CALCULATING $P\{nD_n \leqq c\}$ FOR FINITE n

Lemma. Let $F(x)$ be any continuous cdf and $F_n(x)$ the empirical distribution function resulting from the ordered sample $(x_1, x_2, \ldots, x_n)$. Then the function $F_n(x) - F(x)$ is continuous except for $x \in \{x_1, x_2, \ldots, x_n\}$,[1] where it is discontinuous.

Proof: $F_n(x) - F(x)$ is continuous for all $x \bar{\in} \{x_1, x_2, \ldots, x_n\}$,[1] since the difference of two functions continuous at a point is also continuous at this point.

Consider $x = x_j$ for some $j = 1, 2, \ldots, n$. Then,

$$\lim_{x \to x_j^-} \left[F_n(x) - F(x)\right] = \frac{j-1}{n} - F(x_j)$$

and

$$\lim_{x \to x_j^+} \left[F_n(x) - F(x)\right] = \frac{j}{n} - F(x_j) \quad .$$

---

[1] The notation $x \in \{x_1, x_2, \ldots, x_n\}$ is used to indicate that x "belongs to" the set of points included within the braces. The symbol $\bar{\in}$ is used to indicate "does not belong to".

Hence $F_n(x) - F(x)$ is discontinuous at $x = x_j$ ($j = 1, 2, \ldots, n$) since

$$\lim_{x \to x_j^-} \left[F_n(x) - F(x) \right] \neq \lim_{x \to x_j^+} \left[F_n(x) - F(x) \right] \quad .$$

This completes the proof of the lemma.

Let T_k denote the set of points x such that

$$F(x) = \frac{k}{n} \qquad (k = 1, 2, \ldots, n-1) \quad .$$

Since F is a continuous cdf, T_k ($k = 1, 2, \ldots, n-1$) contains a single point or all the points belonging to some interval of finite length.

Let $g_k = \underset{x \in T_k}{\text{glb}}\ x$ ($k = 1, 2, \ldots, n-1$); and let S denote the set of points $\{g_1, g_2, \ldots, g_{n-1}\}$. Then S is a set of unique points having the properties:

(i) $g_1 < g_2 < \ldots < g_{n-1}$ since F is monotonic non-decreasing;

(ii) for each $g_k \in S$, $F(g_k) = \frac{k}{n}$ since $g_k \in T_k$; and

(iii) $F(x) < \frac{k}{n}$ for all $x < g_k$.

Let c be an integer belonging to the set $\{1, 2, \ldots, n-1\}$. Denote by R_c the set of points x for which the relation

$$F_n(x) - F(x) > \frac{c}{n}$$

holds.

Theorem 3. Every nonempty R_c is bounded above; and if $L = \operatorname{lub}_{x \in R_c} x$ then

$$F_n(L) - F(L) = \frac{c}{n} \quad .$$

This theorem is basic to the development which follows and is given by Feller without proof.

Proof of Theorem 3: We know that

$$\lim_{x \to \infty} F(x) = 1$$

and

$$0 \leqq F(x) \leqq 1$$

since $F(x)$ is a cdf.

There exists a point x_c such that

$$1 - F(x) < \frac{c}{n} \text{ for all } x > x_c \quad . \tag{1}$$

By definition of F_n, we have

$$F_n(x) = 1 \text{ for } x \geqq x_n \quad . \tag{2}$$

Then (1) and (2) imply that

$$F_n(x) - F(x) < \frac{c}{n} \quad \text{for all } x > \max(x_n, x_c) \ .$$

Hence, R_c is bounded above by any $x > \max(x_n, x_c)$.
By the lemma, $F_n(x) - F(x)$ is continuous for $x \in \{x_1, x_2, \ldots, x_n\}$. Suppose that $L = x_j$ for some $j = 1, 2, \ldots, n$.
Since F is continuous and non-decreasing we can choose $\delta > 0$ and points χ_1 and χ_2 such that

$$F(\chi_2) - F(\chi_1) < \frac{1}{n}$$

for

$$x_{j-1} \leq L - \delta < \chi_1 < L < \chi_2 < L + \delta < x_{j+1} \ ^2$$

Since $L = \operatorname{lub}_x R_c$, we can choose the point χ_1 such that it is also true that

$$\chi_1 \in R_c \ .$$

By the definition of F_n we have that

$$F_n(\chi_2) - F_n(\chi_1) = \frac{1}{n} \ .$$

Thus,

---

[2] For $j = 1$ we choose x_{j-1} as any $x < x_1$; and for $j = n$ we choose x_{j+1} as any $x > x_n$.

$$F_n(x_2) - F(x_2) = \frac{1}{n} + F_n(x_1) - F(x_2)$$

$$> \frac{1}{n} + F_n(x_1) - F(x_2) - \frac{1}{n} \quad .$$

That is

$$F_n(x_2) - F(x_2) > F_n(x_1) - F(x_1) > \frac{c}{n}$$

since

$$x_1 \in R_c \quad .$$

This implies that $x_2 \in R_c$, which contradicts $L = \underset{x}{\text{lub}}\, R_c$. Thus the supposition that $L = x_j$ $(j = 1, 2, \ldots, n)$ leads to a contradiction, and we conclude that L must be a continuity point of the function $F_n(x) - F(x)$. Since

$$L = \underset{x}{\text{lub}}\, R_c \quad ,$$

it is true that

$$F_n(x) - F(x) \leqq \frac{c}{n} \qquad \text{for all } x > L$$

and

$$\lim_{x \to L+} \left[F_n(x) - F(x) \right] \leqq \frac{c}{n} \quad .$$

But

$$L = \operatorname{lub}_x R_c$$

also implies that

$$\lim_{x \to L^-} \left[F_n(x) - F(x) \right] = \frac{c}{n} \quad .$$

Finally, since L is a continuity point of $F_n - F$ we have that

$$F_n(L) - F(L) = \frac{c}{n} \quad . \tag{3}$$

This completes the proof of Theorem 3.

Corrolary. $L \in S$

By the definition of F_n, the number $F_n(L)$ must be of the form $\frac{r}{n}$ for some $r \in \{1,2,\ldots,n\}$.

From (3),

$$F(L) = F_n(L) - \frac{c}{n} = \frac{r}{n} - \frac{c}{n}$$

or

$$F(L) = \frac{k}{n} \qquad \text{where } k = r - c \quad .$$

Now $k \neq 0$, for otherwise we would have

$$F(L) = 0 \tag{4}$$

and

$$F(x) = 0 \qquad \text{for all } x \leqq L$$

since F is non-decreasing.

Let $x_2 \in R_c$ (which we have assumed nonempty); then

$$F_n(x_2) > \frac{c}{n} \quad . \tag{5}$$

But

$$x_2 \in R_c \text{ implies that } x_2 < L \quad .$$

Therefore

$$F_n(x_2) \leqq F_n(L) \tag{6}$$

Now (3) and (4) imply

$$F_n(L) = \frac{c}{n}$$

and by (6)

$$F_n(x_2) \leqq \frac{c}{n}$$

which contradicts (5).

Thus $k \neq 0$ and we have that $k \in \{1,2, \ldots, n\}$. Also $k \neq n$, since $r \in \{1, 2, \ldots, n\}$ and $c \in \{1, 2, \ldots, n-1\}$ we have

$$k = r - c \neq n \quad .$$

Therefore

$$k \in \{1, 2, \ldots, n-1\} \quad .$$

We know then that $L \in T_k$ for some k ($k = 1, 2, \ldots, n-1$).

If T_k consists of the single point g_k, it must be true that

$$L = g_k \in S \quad .$$

It remains to consider the case in which T_k contains more than one point. In this case, we need to show that

$$L = \underset{x}{\text{glb}}\ T_k \quad .$$

We know that T_k consists of the points belonging to some interval. Let χ_1 denote the left end point of this interval.

Suppose that

$$L \neq \chi_1 \quad ,$$

that is

$$L > \chi_1 \qquad \text{since } L \in T_k \quad .$$

Let χ_2 be any point in the half-open interval $[\chi_1, L)$. Then

$$F_n(\chi_2) \leq F_n(L) \quad ,$$

by the monotonic property of F_n.

$$F_n(x_2) - F(x_2) \leq F_n(L) - F(L)$$

since

$$F(x_2) = F(L) = \frac{k}{n} \qquad (x_2, L \in T_k) \quad .$$

But,

$$F_n(L) - F(L) = \frac{c}{n} \quad \text{by Theorem 3.}$$

Therefore

$$F_n(x_2) - F(x_2) \leq \frac{c}{n} \quad .$$

This implies that

$$x_2 \bar{\in} R_c \quad .$$

Since the above argument is true for any $x \in [x_1, L)$, we have a contradiction to the fact that $L = \text{lub } R_c$. Thus

$$L = x_1 = \text{glb } T_k$$

$$L = g_k \qquad \text{for some } k = 1, 2, \ldots, n-1$$

or we can say that $L \in S$.

For some $k = 1, 2, \ldots, n-1$ we have from (3) that

$$F_n(g_k) - F(g_k) = \frac{c}{n} \quad .$$

But by the definition of g_k

$$F(g_k) = \frac{k}{n} \quad .$$

Therefore

$$F_n(g_k) = \frac{c+k}{n} \quad .$$

By the definition of F_n we have that

$$x_{k+c} \leq g_k < x_{k+c+1}$$

and in the proof of the lemma we showed that

$$L = g_k * x_j \quad (j = 1, 2, \ldots, n) \quad .$$

Thus

$$x_{k+c} < g_k < x_{k+c+1} \quad .$$

That is, there exists exactly k + c of the X_j's which are smaller than g_k.

Definition 1. Let $A_k(c)$ (k = 1, 2, ..., n-1) denote the event that there are exactly k + c of the X_j's which are smaller than $g_k \in S$.

Theorem 4. For any $c \in \{1, 2, \ldots, n-1\}$ the relation

$$F_n(x) - F(x) > \frac{c}{n}$$

holds for some x (i.e. R_c nonempty) if and only if at least one of the events

$$A_1(c),\ A_2(c),\ \ldots,\ A_{n-1}(c)$$

occurs.

Proof: In the development prior to Definition 1, we have shown that if R_c is nonempty for $c \in \{1, 2, \ldots, n-1\}$ then at least one of the events $A_1(c), A_2(c), \ldots, A_{n-1}(c)$ must occur.

Now suppose $A_k(c)$ occurs for some $k = 1, 2, \ldots, n-1$. Then

$$x_{k+c} < g_k \leqq x_{k+c+1} .$$

Consider a point x_2 such that

$$x_{k+c} \leqq x_2 < g_k \leqq x_{k+c+1}$$

Then

$$F(x_2) < F(g_k) = \frac{k}{n}$$

and

$$F_n(x_2) = F_n(x_{k+c}) = \frac{k+c}{n}$$

Thus we have

$$F_n(x_2) - F(x_2) > \frac{k+c}{n} - \frac{k}{n} = \frac{c}{n}$$

which implies that

$$x_2 \in R_c \quad .$$

This completes the proof of Theorem 4.

In a completely analogous manner, it can be shown that if R_{-c} denotes the set of points for which

$$F_n(x) - F(x) < -\frac{c}{n} \qquad c \in \{1, 2, \ldots, n-1\}$$

and if R_{-c} is nonempty, then there exists

$$G = \text{glb } R_{-c}$$

such that

$$F_n(G) - F(G) = -\frac{c}{n}$$

and

$$G \in S \quad .$$

That is, there exists exactly $k - c$ of the X_j's which are smaller than $g_k \in S$.

Definition 2. Let $A_k(-c)$ $(k = 1, 2, \ldots, n-1)$ denote the event that exactly $k - c$ of the X_j's shall be smaller than $g_k \in S$.

Theorem 5. For any $c \in \{1, 2, \ldots, n-1\}$ the relation $F_n(x) - F(x) < -\frac{c}{n}$ holds for some x (i.e., R_{-c} nonempty) if and only if at least one of the events

$$A_1(-c),\ A_2(-c),\ \ldots,\ A_{n-1}(-c)$$

occurs.

From Theorems 4 and 5 we have

Theorem 6. For any $c \in \{1, 2, \ldots, n-1\}$, the relation

$$\left| F_n(x) - F(x) \right| > \frac{c}{n}$$

holds for some x if and only if at least one of the events

$$A_1(c),\ A_1(-c),\ A_2(c),\ A_2(-c),\ \ldots,\ A_{n-1}(c),\ A_{n-1}(-c)$$

occurs.

We define

$$D_n = \operatorname*{lub}_x \left| F_n(x) - F(x) \right| .$$

For any $c \in \{1, 2, \ldots, n-1\}$, $D_n > \frac{c}{n}$ if $\left| F_n(x) - F(x) \right| > \frac{c}{n}$ for some x. Moreover, by the properties of a least upper bound, $D_n > \frac{c}{n}$ only if $\left| F_n(x) - F(x) \right| > \frac{c}{n}$ for some x. That is,

Theorem 7. For any $c \in \{1, 2, \ldots, n-1\}$, $D_n > \frac{c}{n}$ if and only if at least one of the events

$$A_1(c),\ A_1(-c),\ \ldots,\ A_{n-1}(c),\ A_{n-1}(-c)$$

occurs.

Thus we have

$$P\{nD_n > c\} = P\{A_1(c) + A_1(-c) + \ldots \quad (7)^3$$

$$\ldots + A_{n-1}(c) + A_{n-1}(-c)\}$$

Definition 3. Let U_r ($r = 1, 2, \ldots, n-1$) denote the event that $A_r(c)$ is the first event in the sequence $A_1(c)$, $A_1(-c)$, $A_2(c)$, $\ldots$, $A_{n-1}(c)$, $A_{n-1}(-c)$ to occur.

Definition 4. Let V_r ($r = 1, 2, \ldots, n-1$) denote the event that $A_r(-c)$ is the first event in the sequence $A_1(c)$, $A_1(-c)$, $A_2(c)$, $\ldots$, $A_{n-1}(c)$, $A_{n-1}(-c)$ to occur.

We now show that (7) is equivalent to

$$P\{nD_n > c\} = \sum_{r=1}^{n-1} P\{U_r\} + \sum_{r=1}^{n-1} P\{V_r\} \quad . \qquad (8)$$

By definition

$$U_1 = A_1(c) \quad .$$

$$U_1 + V_1 = A_1(c) + \bar{A}_1(c)A_1(-c) = A_1(c) + A_1(-c)$$

---

[3] The usual notation for sums and products is also used for union and intersection of sets.

Suppose that

$$\sum_{r=1}^{k-1} (U_r + V_r) = \sum_{r=1}^{k-1} \left[A_r(c) + A_r(-c) \right]$$

Then

$$\sum_{r=1}^{k} (U_r + V_r) = \sum_{r=1}^{k-1} (U_r + V_r) + U_k + V_k$$

$$= \sum_{r=1}^{k-1} \left[A_r(c) + A_r(-c) \right] + \bar{A}_1(c)\bar{A}_1(-c)\ldots$$

$$\ldots\bar{A}_{k-1}(-c)A_k(c) + \bar{A}_1(c)\bar{A}_1(-c)\ldots$$

$$\ldots\bar{A}_{k-1}(-c)\bar{A}_k(c)A_k(-c)$$

$$\sum_{r=1}^{k} (U_r + V_r) = \sum_{r=1}^{k-1} \left[A_r(c) + A_r(-c) \right] +$$

$$+ \bar{A}_1(c)\bar{A}_1(-c)\ldots\bar{A}_{k-1}(-c) \left[A_k(c) + \right.$$

$$\left. + \bar{A}_k(c)A_k(-c) \right]$$

$$\sum_{r=1}^{k} (U_r + V_r) = \sum_{r=1}^{k-1} \left[A_r(c) + A_r(-c) \right] + \bar{A}_1(c)\bar{A}_1(-c)\ldots$$

$$\ldots\bar{A}_{k-1}(-c) \left[A_k(c) + A_k(-c) \right]$$

Hence

$$\sum_{r=1}^{k} (U_r + V_r) = \sum_{r=1}^{k} \left[A_r(c) + A_r(-c) \right] .$$

By induction

$$\sum_{r=1}^{n-1} (U_r + V_r) = \sum_{r=1}^{n-1} \left[A_r(c) + A_r(-c) \right] .$$

Since $U_1, U_2, \ldots, U_{n-1}, V_1, V_2, \ldots, V_{n-1}$ are mutually exclusive, we have

$$\sum_{r=1}^{n-1} P\{U_r\} + \sum_{r=1}^{n-1} P\{V_r\} = P\left\{ \sum_{r=1}^{n-1} (U_r + V_r) \right\}$$

$$= P\left\{ \sum_{r=1}^{n-1} \left[A_r(c) + A_r(-c) \right] \right\}$$

$$= P\{nD_n > c\} .$$

That is

$$P\{nD_n > c\} = \sum_{r=1}^{n-1} P\{U_r\} + \sum_{r=1}^{n-1} P\{V_r\} . \qquad (9)$$

Furthermore, the event $A_k(c)$ can occur if and only if one of the events

$$A_k(c)U_1, A_k(c)V_1, \ldots, A_k(c)U_k$$

occurs. Therefore

$$P\{A_k(c)\} = P\left\{ \sum_{r=1}^{k} A_k(c)U_r + \sum_{r=1}^{k-1} A_k(c)V_r \right\} .$$

But these events are mutually exclusive, and

$$P\{A_k(c)\} = \sum_{r=1}^{k} P\{A_k(c)U_r\} + \sum_{r=1}^{k-1} P\{A_k(c)V_r\} .$$

or

$$P\{A_k(c)\} = \sum_{r=1}^{k} P\{U_r\}P\{A_k(c) \mid U_r\} +$$

$$+ \sum_{r=1}^{k-1} P\{V_r\}P\{A_k(c) \mid V_r\} .$$

Since

$$P\{A_k(c) \mid V_k\} = 0$$

we can write

$$P\{A_k(c)\} = \sum_{r=1}^{k} P\{U_r\}P\{A_k(c) \mid U_r\} +$$

$$+ \sum_{r=1}^{k} P\{V_r\}P\{A_k(c) \mid V_r\} .$$

Now, $P\{A_k(c) \mid U_r\}$ is the probability that there are $k + c$

of the X_j's less than g_k under the hypothesis that it is known there are $r + c$ of the X_k's less than g_r, where $r \leqq k$. The given event U_r also implies that $A_r(c)$ is the first event in the sequence

$$A_1(c),\ A_1(-c),\ \ldots,\ A_{n-1}(c),\ A_{n-1}(-c)$$

to occur. Now the random variables $F(X_j)$ are independent and uniformly distributed on the interval $[0,1]$. It follows that $P\{A_k(c) \mid U_r\}$ is equal to the probability that $k + c$ of the X_j's will be less than g_k, given that $r + c$ of the X_j's are less than g_r. That is

$$P\{A_k(c) \mid U_r\} = P\{A_k(c) \mid A_r(c)\} \quad .$$

And similarly

$$P\{A_k(c) \mid V_r\} = P\{A_k(c) \mid A_r(-c)\} \quad .$$

Hence, we have that

$$P\{A_k(c)\} = \sum_{r=1}^{k} P\{U_r\}P\{A_k(c) \mid A_r(c)\} + \tag{10}$$

$$+ \sum_{r=1}^{k} P\{V_r\}P\{A_k(c) \mid A_r(-c)\} \quad ,$$

and

$$P\{A_k(-c)\} = \sum_{r=1}^{k} P\{U_r\}P\{A_k(-c) \mid A_r(c)\} +$$

$$+ \sum_{r=1}^{k} P\{V_r\}P\{A_k(-c) \mid A_r(-c)\} \quad .$$

These last two relations hold for k = 1, 2, ..., n-1.

The system (10) of 2n-2 equations in the 2n-2 unknowns $P\{U_r\}$, $P\{V_r\}$ (r = 1, 2, ..., n-1) is essentially given by Feller from which he derives the limiting distribution of Kolmogorov and Smirnov. In the following paragraphs this system is reduced to a set of recursive relations from which $P\{U_r\}$ and $P\{V_r\}$ can be calculated. These values used in (9) then give the exact probability, $P\{nD_n > c\}$, for finite n.

From the definitions of the events $A_k(c)$ and $A_k(-c)$ we have the following relations which make it possible to reduce the system of equations (10) to a set of recursive relations for calculating $P\{U_r\}$ and $P\{V_r\}$:

$$\text{(a)} \quad P\{A_k(c) \mid A_r(c)\} = 0 \quad \text{for } k > n - c \quad .$$

$A_k(c)$ is the event that k + c of the X_j's will be smaller than the given number g_k. This event obviously cannot occur if k > n - c since there are only n of the X_j's. From the definition of conditional probability we have,

$$P\{A_k(c) \mid A_r(c)\} = 1 \quad \text{for } k = r$$

$$(b)\quad P\{A_k(c) \mid A_r(-c)\} = 0 \quad \text{for } r = 1, 2, \ldots, c-1$$

$$k = r,$$

$$k > n-c \quad .$$

For $r = 1, 2, \ldots, c-1$ we have that $r - c < 0$ and hence the event $A_r(-c)$ cannot occur. For $k = r$, the events $A_k(c)$ and $A_k(-c)$ cannot both occur since $c \neq 0$. As in (a) above, $A_k(c)$ cannot occur if $k > n-c$.

$$(c)\quad P\{A_k(-c) \mid A_r(c)\} = 0 \quad \text{for } k = r,$$

$$k = 1, 2, \ldots, c-1,$$

$$r > k - 2c \quad .$$

For $k = r$ or $k = 1, 2, \ldots, c-1$ the reasoning is the same as before. Consider $r > k - 2c$. If the event $A_r(c)$ occurs, then there are exactly $r + c$ of the X_j's less than g_r. For $A_k(-c)$ to occur, $k - c$ of the X_j's must be less than g_k. Thus for both events to occur it must be true that there are

$$k - c - r - c = k - 2c - r$$

of the X_j s in the interval $g_r, g_k)$. For $r > k - 2c$, however,

$$k - 2c - r < 0$$

and thus

$$P\{A_k(-c) \mid A_r(c)\} = 0 \qquad \text{for } r > k - 2c$$

In an analogous manner we have,

$$\text{(d)} \quad P\{A_k(-c) \mid A_r(-c)\} = 0 \qquad \text{for } k < c$$

$$r < c$$

$$= 1 \qquad \text{for } r = k$$

$$\text{(e)} \quad P\{A_k(c)\} = 0 \qquad \text{for } k > n - c$$

$$\text{(f)} \quad P\{A_k(-c)\} = 0 \qquad \text{for } k < c \quad .$$

To simplify notation, let

$$a_{kr} = P\{A_k(c) \mid A_r(c)\}$$

$$b_{kr} = P\{A_k(c) \mid A_r(-c)\}$$

$$c_{kr} = P\{A_k(-c) \mid A_r(c)\}$$

$$d_{kr} = P\{A_k(-c) \mid A_r(-c)\}$$

$$s_k = P\{A_k(c)\}$$

$$t_k = P\{A_k(-c)\}$$

$$u_k = P\{U_k\}$$

$$v_k = P\{V_k\} \quad .$$

Then (9) and (10) can be written as,

$$P\{nD_n > c\} = \sum_{k=1}^{n-1} u_k + \sum_{k=1}^{n-1} v_k \tag{11}$$

and

$$s_k = \sum_{r=1}^{k} u_r a_{kr} + \sum_{r=1}^{k} v_r b_{kr} \tag{12}$$

$$t_k = \sum_{r=1}^{k} u_r c_{kr} + \sum_{r=1}^{k} v_r d_{kr}$$

for k = 1, 2, ..., n-1. Considering the relations (a) through (f) above, these reduce to:

$$P\{nD_n > c\} = \sum_{k=1}^{n-c} u_r + \sum_{k=c}^{n-1} v_r \tag{13}$$

and

$$s_1 = u_1 \tag{14}$$

$$s_k = u_k + \sum_{r=1}^{k-1} u_r a_{kr} \qquad (k = 2, 3, \ldots, c)$$

$$s_k = u_k + \sum_{r=1}^{k-1} u_r a_{kr} + \sum_{r=c}^{k-1} v_r b_{kr} \qquad (k = c+1, \ldots, n-c)$$

$$t_c = v_c$$

$$t_k = v_k + \sum_{r=c}^{k-1} v_r d_{kr} \qquad (k = c+1, \ldots, 2c)$$

$$t_k = v_k + \sum_{r=c}^{k-1} v_r d_{kr} + \sum_{r=1}^{k-2c} u_r c_{kr} \qquad (k = 2c+1, \ldots, n-1)$$

Solving (14) for u_k and v_k:

$$u_1 = s_1 \tag{15}$$

$$u_k = s_k - \sum_{r=1}^{k-1} u_r a_{kr} \qquad (k = 2, \ldots, c)$$

$$u_k = s_k - \sum_{r=1}^{k-1} u_r a_{kr} - \sum_{r=c}^{k-1} v_r b_{kr} \qquad (k = c+1, \ldots, n-c)$$

$$v_c = t_c$$

$$v_k = t_k - \sum_{r=c}^{k-1} v_r d_{kr} \qquad (k = c+1, \ldots, 2c)$$

$$v_k = t_k - \sum_{r=c}^{k-1} v_r d_{kr} - \sum_{r=1}^{k-2c} u_r c_{kr} \qquad (k = 2c+1, \ldots, n-1)$$

from which u_k ($k = 1, 2, \ldots, n-c$) and v_k ($k = c, c+1, \ldots$ $\ldots, n-1$) can be calculated once a_{kr}, b_{kr}, c_{kr}, d_{kr}, s_k, t_k are known. Feller has shown that these quantities are

certain binomial probabilities.

By the definition of g_k,

$$P\{X_j \leq g_k\} = F(g_k) = \frac{k}{n} \quad .$$

Thus

$$s_k = \binom{n}{k+c} \left(\frac{k}{n}\right)^{k+c} \left(1 - \frac{k}{n}\right)^{n-k-c}$$

since $A_k(c)$ is the event that exactly $k + c$ of the variables X_j shall be less than g_k.

Similarly

$$t_k = \binom{n}{k+c} \left(\frac{k}{n}\right)^{k-c} \left(1 - \frac{k}{n}\right)^{n-k+c} \quad .$$

Now,

$$P\{A_k(c) \mid A_r(c)\} = \frac{P\{A_k(c)A_r(c)\}}{P\{A_r(c)\}}$$

It is easily shown that the expression on the right of the above equation is

$$\frac{\frac{n!}{(r+c)!\,(k-r)!\,(n-k-c)!}\left(\frac{r}{n}\right)^{r+c}\left(\frac{k-r}{n}\right)^{k-r}\left(1-\frac{k}{n}\right)^{n-k-c}}{\frac{n!}{(r+c)!\,(n-r-c)!}\left(\frac{r}{n}\right)^{r+c}\left(\frac{n-r}{n}\right)^{n-r-c}}.$$

Reducing the above expression we finally obtain

$$a_{kr} = \binom{n-r-c}{k-r}\left(\frac{k-r}{n-r}\right)^{k-r}\left(1-\frac{k-r}{n-r}\right)^{n-k-c}$$

In like manner, we have

$$b_{kr} = \binom{n-r+c}{k-r+2c}\left(\frac{k-r}{n-r}\right)^{k-r+2c}\left(1-\frac{k-r}{n-r}\right)^{n-k-c}$$

$$c_{kr} = \binom{n-r-c}{k-r-2c}\left(\frac{k-r}{n-r}\right)^{k-r-2c}\left(1-\frac{k-r}{n-r}\right)^{n-k+c}$$

$$d_{kr} = \binom{n-r+c}{k-r}\left(\frac{k-r}{n-r}\right)^{k-r}\left(1-\frac{k-r}{n-r}\right)^{n-k+c}$$

Thus, the probability $P\{nD_n > c\}$ can be calculated by (13) for given integers n and c.

CHAPTER III

CALCULATION OF $P\{nD_n \leqq c\}$ FOR CERTAIN FINITE n

The relations (13) and (15) derived in Chapter II provide a method by which the probability that nD_n shall exceed c can be calculated given positive integers n and c, $(c < n)$. To illustrate these results, Table 1 has been established. Table 1 shows $P\{nD_n \leqq c\}$ for

$$n = 2,3,\ldots,19,20;25;30;35;40;45;50;100$$

and

$$c = 1,2,\ldots,n-1 \quad \text{for each } n \quad .$$

The relation

$$P\{nD_n \leqq c\} = 1 - P\{nD_n > c\}$$

has been used where $P\{nD_n > c\}$ was calculated by relations (13) and (15). The calculations for n through twenty were accomplished by use of the IBM 650 Magnetic Drum Data-Processing Machine, and for n greater than twenty by use of the Remington Rand 1105 Scientific Electronic Computer. The program magnetic tapes for these calculations are on file at the Research Computation Center, University of North Carolina

and can be used to find $P\{nD_n \leqq c\}$ for any specified c and n. The results given in Table 1 agree completely with similar tables first calculated by Massey [9] and corrected and extended by Birnbaum [10].

Values of the limiting distribution, L(z), have been given by Smirnov [11]. The difference between the value of the limit distribution and the exact distribution can be obtained by the following method.

From Kolmogorov s theorem we have that

$$P\{\sqrt{n}\, D_n \leqq z\} \longrightarrow L(z) \quad \text{for } z > 0$$

as $n \longrightarrow \infty$. Now for $c = z\sqrt{n}$,

$$P\{nD_n \leqq c\} = P\{\sqrt{n}\, D_n \leqq z\} \quad .$$

Thus for $z\sqrt{n}$ an integer, relations (13) and (15) can be used to find the exact probability corresponding to the limit distribution, L(z). In Table 2 the difference

$$P\{nD_n \leqq c\} - L(z)$$

is given for certain values of z and n. The entries in this table have been formed again by direct use of (13) and (15).

Table 1. $P\{nD_n \leq c\}$

c \ n	2	3	4	5	6	7
1	.500000	.200000	.093751	.038400	.015432	.006120
2		.925926	.812500	.691200	.576500	.474462
3			.992188	.969920	.934414	.889374
4				.999360	.996228	.989113
5					.999957	.996041
6						.999998

Table 1. (Cont.)

n / c	8	9	10	11	12	13
1	.002403	.000937	.000363	.000140	.000052	.000021
2	.386591	.312614	.251281	.201002	.160136	.127148
3	.838424	.784423	.729464	.675019	.622193	.571357
4	.977409	.961201	.941011	.917465	.891261	.863041
5	.998491	.996151	.992223	.986485	.978853	.969350
6	.999964	.999818	.998651	.997320	.995304	.992497
7	1	.999997	.999980	.999925	.999793	.999529
8		1	1	.999998	.999991	.999971
9				1	1	.999999
10						1

Table 1. (Cont.)

n / c	14	15	16	17	18	19
1	.000007	.000003	.000001	0	0	0
2	.100664	.079497	.062645	.049270	.038687	.030329
3	.523228	.477950	.435635	.396313	.359909	.326359
4	.833375	.802750	.771575	.740189	.708870	.677840
5	.958074	.945170	.930808	.915171	.898440	.880786
6	.992497	.988825	.984247	.978751	.972346	.965059
7	.999176	.998371	.997355	.995975	.994186	.991951
8	.999925	.999836	.999680	.999536	.999172	.998563
9	.999996	.999989	.999973	.999943	.999890	.999805
10	1	1	.999999	.999996	.999991	.999980
11			1	1	.999999	.999999
12					1	1

Table 1. (Cont.)

c \ n	20	25	30	35	40	45
1	0	0	0	0	0	0
2	.023745	.006871	.001948	.000545	.000151	.000041
3	.295533	.177023	.103922	.061636	.034477	.019603
4	.647280	.505537	.386931	.292047	.218189	.161741
5	.862374	.763679	.662889	.567441	.480779	.404177
6	.956993	.905645	.842031	.772497	.701586	.632235
7	.989245	.968323	.935876	.894473	.847071	.796226
8	.997881	.991096	.977445	.956638	.929519	.897390
9	.999676	.997918	.993148	.984234	.970768	.952927
10	.999962	.999598	.998209	.994940	.989104	.980335
11	.999997	.999937	.999600	.998570	.996357	.992528
12	1	.999992	.999924	.999646	.998910	.997421
13		.999999	.999988	.999923	.999709	.999193
14		1	.999998	.999986	.999931	.999771
15			1	.999998	.999985	.999942
16				1	.999997	.999987
17					1	.999997
18						.999999
19						1

Table 1. (Cont.)

c \ n	50	100
1	0	0
2	.000011	
3	.011078	.000031
4	.119160	.004692
5	.337689	.046784
6	.566232	.157115
7	.743923	.315329
8	.861604	.481781
9	.931224	.629368
10	.968561	.747307
11	.986790	.835037
12	.994903	.896696
13	.998196	.937908
14	.999415	.964175

c \ n	50	100
15	.999826	.980161
16	.999953	.989457
17	.999988	.994624
18	.999997	.997370
19	.999999	.998767
20	1	.999554
21		.999445
22		.999761
23		.999894
24		.999985
25		.999995
26		.999998
27		.999999
28		1

Table 2. $P\{nD_n \leqq c\} - L(z)$

z	.30	.33	.40	.50	.60	.67
L(z)	.000009	.000091	.002808	.036055	.135718	.239582
4				.057696		
9		.000846				.073032
16				.026590		
25			.004063		.031305	
36		.000331		.017808		.036122
49						
64				.013400		
81		.000205				.022952
100	.000022		.001884	.010729	.021397	

Table 2. (Cont.)

z	.75	.80	1.00	1.50	2.00	2.50
L(z)	.372833	.455857	.730000	.977782	.999329	.9999925
n						
4			.051250	.014406	.000671	.0000075
9			.054423		.000479	
16	.062802		.041575	.006465	.000351	.0000065
25		.049680	.033679		.000269	
36			.028311	.004124	.000215	.0000035
49			.024424		.000178	
64	.033153		.021478	.003019	.000152	.0000025
81			.019168		.000131	
100		.025924	.017307	.002379	.000125	.0000025

Literature Cited

[1] R. von Mises, "Wahrscheinlichkeitsrechung", F. Deuticke and Wien, (1931), p. 316.

[2] A. Kolmogorov, "Confidence Limits for an Unknown Distribution Function", Annals of Mathematical Statistics, Vol. 12, (1941), pp. 461-463.

[3] N. Smirnov, "Ob Uklonenijah Empiriiceskoi Krivor Raspredelenija", Recueil Mathematique, N.S.Vol. 6(48), (1939), pp. 3-26.

[4] A. Kolmogorov, "Sulla Determinazione Empirica di una Legga di Distribuzione", Istituto Italiano deglia Attuari, Vol. 4, (1933), pp. 1-11.

[5] N. Smirnov, "On the Estimation of the Discrepancy Between Empirical Curves of Distribution for Two Independent Samples", Bulletin Mathematique de l'Universite de Moscow, Vol. 2, (1939), fasc. 2.

[6] W. Feller, "On the Kolmogorov-Smirnov Limit Theorems for Empirical Distributions", Annals of Mathematical Statistics, Vol. 19, (1948), pp. 177-190.

[7] J.L. Doob, "Heuristic Approach to the Kolmogorov-Smirnov Theorems", Annals of Mathematical Statistics, Vol. 20, (1949), pp. 393-403.

[8] K. Chung, "An Estimate Concerning the Kolmogorov Limit Distribution", Transactions of the American Mathematical Society, Vol. 67, (1949), pp. 36-50.

[9] K.J. Massey, "A Note on the Estimation of a Distribution Function by Confidence Limits", Annals of Mathematical Statistics, Vol. 21, (1950), pp. 116-119.

[10] Z.W. Birnbaum, "Numerical Tabulation of the Distribution of Kolmogorov's Statistics for Finite Sample Size", Journal of the American Statistical Association, Vol. 47, (1952), pp. 435-441.

[11] N. Smirnov, "Table for Estimating the Goodness of Fit of Empirical Distributions", Annals of Mathematical Statistics, Vol. 19, (1948), pp.279-281.

Other References

Anderson, T.W. and D.A. Darling, "Asymptotic Theory of Certain 'Goodness of Fit' Criteria Based on Stochastic Processes", Annals of Mathematical Statistics, Vol. 23, (1952), pp. 193-212.

Anderson, T.W. and D.A. Darling, "A Test of Goodness of Fit", Journal of the American Statistical Association, Vol. 49, (1954), pp. 765-769.

Blackman,J., "On the Approximation of a Distribution Function by an Empirical Distribution", Annals of Mathematical Statistics, Vol. 26, (1955), pp. 256-267.

Dansher, M.D., "Justification and Extension of Doob's Heuristic Approach to the Kolmogorov-Smirnov Theorems", Annals of Mathematical Statistics, Vol. 23, (1952), pp. 271-281.

Massey, F.J., "The Distribution of the Maximum Deviations Between Two Sample Cumulative Step Functions", Annals of Mathematical Statistics, Vol. 22, (1951), pp. 125-128.

Wald, A. and Wolfowitz, "Confidence Limits for Continuous Distribution Functions", Annals of Mathematical Statistics, Vol. 10, (1939), pp. 105-118.